APPENDIX B

Journal of Structural Biology 134, 204–218 (2001)
doi:10.1006/jsbi.2000.4336, available online at http://www.idealibrary.com on IDEAL®

Review: Protein Secondary Structure Prediction Continues to Rise

Burkhard Rost

CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, New York 10032

Received November 20, 2000, and in revised form February 21, 2001; published online June 13, 2001

Methods predicting protein secondary structure improved substantially in the 1990s through the use of evolutionary information taken from the divergence of proteins in the same structural family. Recently, the evolutionary information resulting from improved searches and larger databases has again boosted prediction accuracy by more than four percentage points to its current height of around 76% of all residues predicted correctly in one of the three states, helix, strand, and other. The past year also brought successful new concepts to the field. These new methods may be particularly interesting in light of the improvements achieved through simple combining of existing methods. Divergent evolutionary profiles contain enough information not only to substantially improve prediction accuracy, but also to correctly predict long stretches of identical residues observed in alternative secondary structure states depending on nonlocal conditions. An example is a method automatically identifying structural switches and thus finding a remarkable connection between predicted secondary structure and aspects of function. Secondary structure predictions are increasingly becoming the work horse for numerous methods aimed at predicting protein structure and function. Is the recent increase in accuracy significant enough to make predictions even more useful? Because the recent improvement yields a better prediction of segments, and in particular of β strands, I believe the answer is affirmative. What is the limit of prediction accuracy? We shall see.

INTRODUCTION

History. Linus Pauling correctly guessed the formation of helices and strands (14, 15) (and falsely hypothesized other structures). Three years before Pauling's guess was verified by the publications of the first X-ray structures (16, 17), one group had already ventured to predict secondary structure from sequence (18). The first-generation prediction methods following in the 1960s and 1970s were all based on single amino acid propensities (19). The second-generation methods dominating the scene until the early 1990s used propensities for segments of 3–51 adjacent residues (19). Basically any imaginable theoretical algorithm had been applied to the problem of predicting secondary structure from sequence. However, it seemed that prediction accuracy stalled at levels slightly above 60% (percentage of residues predicted correctly in one of the three states: helix, strand, and other). The reason for this limit was the restriction to local information. Can we introduce some global information into local stretches of residues?

Secondary structure prediction profits from divergence. Early on, Dickerson *et al.* (20) realized that information contained in multiple alignments can improve predictions. Zvelebil *et al.* (21) incorporated this concept into an automatic prediction method. However, the breakthrough of the third-generation methods to levels above 70% accuracy required a combination of larger databases with more advanced algorithms (19, 22). The major component of these new methods was the use of evolutionary information. All naturally evolved proteins with more than 35% pairwise identical residues over more than 100 aligned residues have similar structures (23). This seemingly implies an amazing stability of structure with respect to sequence divergence. However, this average figure hides the fact that neutral mutations are extremely unlikely. Supposedly most mutations result in proteins that will not adopt any globular structure, at all. In other words, only a tiny fraction of all possible proteins exist. Hence, position-specific profiles describing which residues can be exchanged against which others at which positions contain crucial information about protein structure. One consequence is that stretches of say 17 adjacent residues implictly contain some information about long-range interactions and environment since the profile reflects evolutionary constraints. Using evolutionary divergence was the start key to the third-generation prediction meth-

1047-8477/01 $35.00

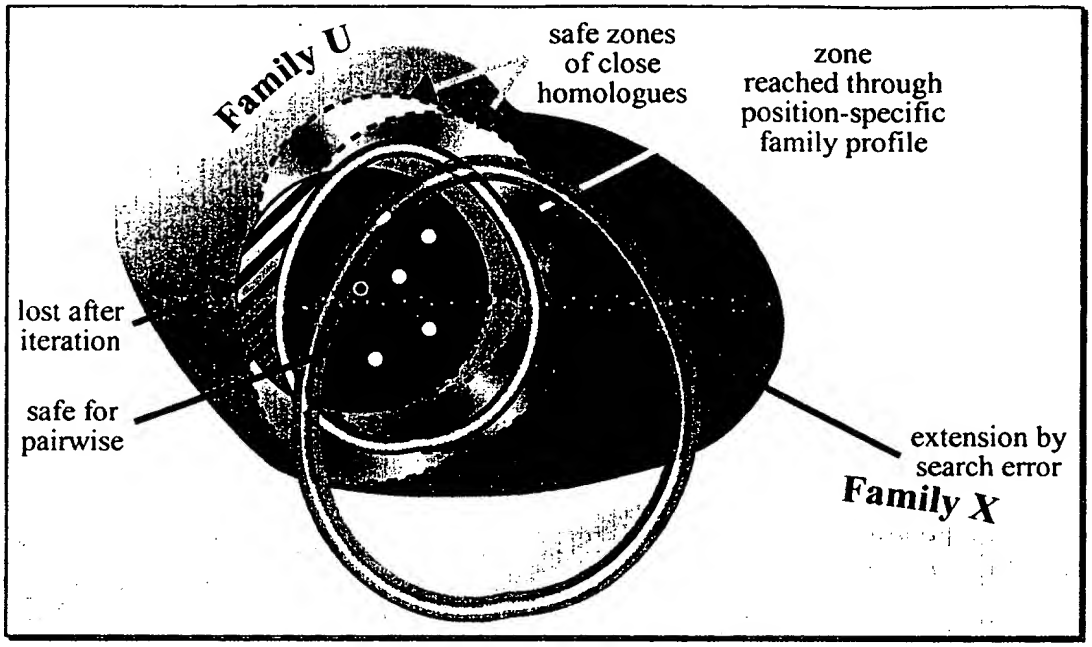

FIG. 1. Profile-based searches extend evolutionary information. The cloud signifies a protein structural family for the query protein U, i.e., all proteins that have a similar 3D structure. A simple pairwise comparison of U with all other proteins covers the "safe zone" of sequence alignment (gray circle around U). This zone can be defined, e.g., by BLAST scores below 10^{-10} or by more than 35% pairwise identical residues over long alignments. Assume that there are only five other proteins (small white circles) in the safe zone falling on one side of U. For example, PSI-BLAST starts the next iteration with the family-specific profile given by the proteins found in the safe zone. Searching the database again with this profile reaches safely into the twilight zone (zone reached marked by double-lined egg indicated in figure). However, no current method generally reaches all members of family U. Furthermore, in particular for PSI-BLAST the new region may fall outside of the initial safe zone (black subregion of the safe zone). Finally, the regions that could have been reached by sequence-space hopping or intermediate sequence searches (dashed circles around five initial hits; (120, 121)) are not entirely covered by the profile-based search. The tricky bit is to avoid the possibility that the profile will pick unrelated proteins (transparent egg) and thus connect two separate structural families (U and X). *Conclusions:* (i) Iterated PSI-BLAST searches can safely identify fairly divergent family members. (ii) Close homologues may be lost during the extension of the family. (iii) The advanced search can lead the results astray.

ods. Knowing 3D structure,[1] we can identify very distant relationships between proteins that would improve accuracy even further (24). Can we build larger and more diverged families without knowing structure?

[1] Abbreviations used: 3D structure, three-dimensional (coordinates of protein structure); 1D structure, one-dimensional (e.g., sequence or string of secondary structure); ASP, method identifying regions of structure ambivalent in response to global changes (1); DSSP, database and method converting 3D coordinates into secondary structure (2); HMMSTR, hidden Markov model-based prediction of secondary structure (3); JPred, method combining other prediction methods (4, 5); JPred2, divergent profile (PSI-BLAST)-based neural network prediction (6); PHD, simple profile-based neural network prediction (7); PHDpsi, divergent profile (PSI-BLAST)-based neural network prediction (7, 8); PROF, divergent profile-based neural network prediction trained and tested with PSI-BLAST (9); PSI-BLAST, gapped and iterative specific profile-based, fast and accurate alignment method (10); PSIPRED, divergent profile (PSI-BLAST)-based neural network prediction (11); SAM-T99sec, neural network prediction, using hidden Markov models as input (12); SSpro, profile-based advanced neural network prediction method (13).

New database searches extend family divergence. It was also recognized very early on that information from the position-specific evolutionary exchange profile of a particular protein family facilitates discovering more distant members of that family (20). Automatic database search methods successfully used position-specific profiles for searching (25). However, the breakthrough for large-scale routine searches was achieved with the development of PSI-BLAST (10) and hidden Markov models (12, 26). In particular, the gapped, profile-based, and iterated search tool PSI-BLAST continues to revolutionize the field of protein sequence analysis through its unique combination of speed and accuracy. More distant relationships are found through iteration starting from the safe zone of comparisons and intruding deeply and reliably into the twilight zone (Fig. 1).

Topics left out here. This review focuses on methods predicting secondary structure for globular proteins, in general. At the infancy of analyzing the

proteome of entirely sequenced organisms, the most useful structure prediction methods are those that focus on particular classes of proteins, such as proteins containing membrane helices and coiled-coil regions (27–30). For predicting the topology of helical membrane proteins, a number of new methods add interesting new facets (31–36). However, no method has truly used the flood of recent experimental information about membrane proteins (37). Overall, membrane helices can be predicted much more accurately than globular helices. The current state of the art is to correctly predict all membrane helix topology for more than 80% of the proteins and to falsely predict membrane helices for less than 4% of all globular proteins. We have recently come across evidence suggesting that this figure overestimates performance (Rost, unpublished). Clearly, methods developed to predict helices in globular proteins go completely wrong for membrane helices! In contrast, porins appear to be predicted relatively accurately by methods developed for globular proteins (38, 39). Few methods specifically predicting coiled-coil regions have been published recently (older review in (40)). Two interesting developments are the prediction of the dimeric state of coiled-coils (41) and a method predicting 3D structure for coiled-coil regions (42). In fact, the latter is the only existing method predicting 3D structure below 2-Å main chain deviation over more than 30 residues. Another example of successful specialized secondary structure prediction methods is the focus on β turns (43, 44). The method from the Thornton group appears to be the most accurate current means of predicting turns. Successful methods specialized in predicting α-helix propensities have resulted from the experimental studies of short peptides in solution (45, 46). Neither the turn nor the helix-in-solution methods have yet been combined with other secondary structure prediction methods.

MORE DATA + REFINED SEARCH = BETTER PREDICTION

Jones broke through by using PSI-BLAST searches of large databases. David Jones pioneered the use of iterated PSI-BLAST searches automatically (11). The most important step achieved by the resulting method PSIPRED has been the detailed strategy of avoiding pollution of the profile through unrelated proteins (Fig. 1). To avoid this trap, the database searched must be filtered first (11). At the CASP meeting at which David Jones introduced PSIPRED, Kevin Karplus and colleagues presented their prediction method (SAM-T99sec), finding more diverged profiles through hidden Markov models (47, 48). Recently, Cuff and Barton also successfully used PSI-BLAST alignments for JPred2 (see 49). Jennings *et al.* (50) explore an alternative to increasing divergence: they started with a safe zone alignment through ClustalW (51) and HMMer (26) and iteratively refined the alignment using the secondary structure prediction from DSC (52). The resulting alignment is reported to be more accurate and to yield higher prediction accuracy than the initial ClustalW/HMMer alignments (50). How accurate is secondary structure prediction in 2000?

Prediction accuracy peaks at 76% accuracy. The current best methods reach a level of 76% three-state per-residue accuracy (Table I). This constitutes a sustained level more than four percentage points above the last century's best method not using diverged profiles (PHD in Table I). Fortunately, the improvement is valid for helix, strand, and nonregular regions (information and correlation indices in Table I). Furthermore, significantly fewer residues are confused between the states helix and strand (BAD score, Table I). Finally, some new methods also improve in a more global sense by improving the accuracy of assigning the secondary structural class (all-alpha, all-beta, alpha/beta, and other) based on the predicted content of regular secondary structure (Class score, Table I).

Sources of improvement: Four parts database growth, three parts extended search, two parts other. Jones solicited two causes for the improved accuracy: (i) training and (ii) testing the method on PSI-BLAST profiles. Cuff and Barton examined in detail how different alignment methods improve (6). However, which fraction of the improvement results from the mere growth of the database, which fraction results from using more diverged profiles, and which fraction results from training on larger profiles? Using PHD from 1994 to separate the effects (8), we first compared a noniterative standard BLAST (53) search against SWISS-PROT (54) with one against SWISS-PROT + TrEMBL (54) + PDB (55). The larger database improves performance by about two percentage points (8). Second, we compared the standard BLAST against the large database with an iterative PSI-BLAST search. This yielded less than two percentage points in additional improvement (8). Thus, overall, the more divergent profile search against today's databases supposedly improves any method using alignment information by almost four percentage points (PHDpsi in Table I). The improvement gained by using PSI-BLAST profiles to develop the method is relatively small: PHDpsi was trained on a small database of not very divergent profiles in 1994; e.g., PROF was trained on PSI-BLAST profiles of a 20 times larger database in 2000. The two differ by only one percentage point (Table I), and part of

TABLE I

Accuracy of Secondary Structure Prediction Methods[a]

Method[b]	Q_3[c]	Q_3 Claim[d]	SOV[e]	Info[f]	CorrH[g]	CorrE[h]	CorrL[i]	Class[k]	BAD[l]
PROF	77.0		73	0.37	0.67	0.65	0.56	82	2.2
PSIPRED	76.6	76.5–78.3[m]	73	0.37	0.66	0.64	0.56	81	2.5
SSpro	76.3	76	71	0.36	0.67	0.64	0.56	83	2.5
JPred2	75.2	76.4	70	0.34	0.65	0.63	0.54	77	2.4
PHDpsi	75.1		70	0.29	0.64	0.62	0.53	80	2.9
PHD	71.9	71.6	68	0.25	0.59	0.59	0.49	77	4.1
Copenhagen	78[n]	77.8							
Wang/Yuan								53[o]	

[a] Data set and sorting: The results are compiled by EVA (58). All methods for which details are listed have been tested on 195 different new protein structures (EVA version February 2001). None of these proteins was similar to any protein used to develop the respective method. This set comprised the largest such set by February 1, 2001, for which we had results. Sorting and grouping reflect the following concept: if the data set is too small to distinguish between two methods, these two are grouped. For the given set of 195 proteins, this yielded three groups. Inside of each group, results are sorted alphabetically. Due to a lack of data, I could not add the performance of SAM-T99sec (48); on a set of 105 proteins SAM-T99sec appears comparable to the best three methods: PSIPRED, SSpro, and PROF. The results from the Copenhagen method are set apart, since they were not collected continuously by EVA (the method is not publicly available); rather they were provided by the group in Denmark for this review and thus may have been based on marginally differing sequence databases.

[b] See abbreviations footnote in text; Copenhagen refers to the method from the group in Denmark (63); Wang/Yuan refers to a method predicting secondary structural class from the amino acid composition, which may be the most accurate such method (59).

[c] Three-state per-residue accuracy, i.e., number of residues predicted correctly in one of the three states, helix, strand, or other (conversion of DSSP states (HG) → helix, (EB) → strand; note that the per-residue accuracy tends to favour methods overpredicting nonregular structure).

[d] Three-state per-residue accuracy published in original publication of method: PSIPRED (11), SSpro (13), JPred2 (6), PHD (122).

[e] Three-state per-segment score measuring the overlap between predicted and observed segments (75, 123).

[f] Per-residue information content (22).

[g] Matthew's correlation coefficient for state helix (124).

[h] Matthew's correlation for state strand (124).

[i] Matthew's correlation coefficient for state other (124).

[k] Percentage of proteins correctly sorted into one of the four classes: all-alpha (length > 60, helix >45%, strand <5%), all-beta (length > 60, helix <5%, strand >45%), alpha/beta (length > 60, helix >30%, strand >20%), other (thresholds for classification from (122, 125, 126).

[l] Percentage of helical residues predicted as strand and of strand residues predicted as helix (127).

[m] PSIPRED results were published for different conversions of the eight DSSP states to three states.

[n] P.

[o] The class accuracy for the method based on amino acid composition is taken from the original publication (59), i.e., based on a different data set than all other methods.

this difference resulted from implementing new concepts into PROF (Rost, unpublished; 9).

CAUTION: OVEROPTIMISM HAS BECOME EVEN MORE LIKELY![2]

Seemingly improving accuracy by ignoring short segments. There are many ways to publish higher levels of accuracy. Among the simplest for secondary structure prediction is to convert 3_{10} helices and β bulges assigned by DSSP (2) to nonregular structure. This yields higher levels of accuracy since all methods—on average—are better at predicting the middle of helices and strands than their caps and hence are more accurate for longer regular secondary structure segments (56, 57). When predicted secondary structure is used to predict 3D structure, short helices are important. Thus, I suggest bearing with the more conservative conversion strategy.

Comparing apples and oranges or too few apples with one another. To overstate the point: there is NO value in comparing methods evaluated on different data sets. Most secondary structure prediction methods are available. Thus, developers may want to compare their results to public methods based on the same data set (not previously used for either of the two). Many methods predicting aspects of protein structure and function must fight with limited data availability. This is not at all the case for secondary structure prediction. Hundreds of new protein structures are added every year (55). If for some reason or another, small data sets must be used, developers should painstakingly try to estimate what "significant difference" means for their data set. For example, 16 new protein structures are clearly too few! We currently have results from

[2] Note: I added this section listing "what not-to do" primarily for developers of methods, since many of the recently published methods fall prey to one of the problems mentioned.

many prediction methods for 16 proteins. For that set, JPred2, PHD, PROF, PSIPRED, SAM-T99sec, and SSpro are indistinguishable (58)!

Seemingly achieve 100% accuracy by using correlated sets. Many publications on predicting secondary structural class from amino acid composition allowed correlations between "training" and testing sets. Consequently, levels of prediction accuracy published far exceeded the possible theoretical margins (59). A very simple operational definition for "independent sets" is the following: Two proteins A and B are correlated if the sequence similarity between A and B suffices to predict the structure of B knowing A's structure. Assume we have two uncorrelated sets of proteins S1 and S2. Can we train the method on set S1 and develop it on set S2 without further ado? While developing PROF, I realized that the answer is negative. In fact, I trained neural networks on about 2000 structures that had no significant level of sequence similarity to our original set of 126 proteins (22). I used the 126 proteins only after I had completed developing the method and found a prediction accuracy exceeding 80% (unpublished). When I tested PROF on a set of about 200 new structures that had been added to PDB in the meantime (different from that given in Table I), prediction accuracy dropped. Do the 126 proteins differ from the set used for Table I? I failed to answer this question. Conclusion: test as test can; i.e., use as many independent sets of new structures as possible!

EVA: Automatic evaluation of automatic prediction servers. In collaboration with Volker Eyrich (Columbia), Marc Marti-Renom and Andrej Sali (both from Rockefeller), and Florencio Pazos and Alfonso Valencia (both from CNB Madrid), we have started to address the above problems through the automatic server EVA (58). Leszek Rychlewski (IIMCB Warsaw) and Dani Fischer (Ben-Gurion University) are implementing similar ideas in LiveBench (60). The simple concept is the following: Take the N newest experimental structures added to PDB, send the sequences to all prediction servers, collect the results, and accumulate a continuous evaluation of prediction accuracy every week. EVA has been evaluating secondary structure prediction methods for more than 6 months now. I found it instructive to see how the "ranking" of methods initially changed from week to week due to too small sets. Currently, EVA also provides results for evaluating comparative modeling (Sali group) and residue–residue contacts (Valencia group). We hope that EVA will eventually simplify life for developers, referees, editors, and users.

CLEVER METHODS CAN BE MORE ACCURATE

SSpro: Advanced recursive neural network system. The only method published recently that appears to improve prediction accuracy significantly not through more divergent profiles but through the particular algorithm is SSpro (13). The major idea of the method aims at solving the following problem. When, e.g., training neural networks it is important to avoid correlations between training samples presented successively to the system. A neural network may be presented with the window around residue 11 in protein X at time step T and residue 7 in protein Y at step $T + 1$. Thus, the system never learns that secondary structure correlates between adjacent residues. The result is that regular secondary structure segments are predicted—on average—at a length half that observed (19). PHD addressed this problem by a second-level structure-to-structure network that was trained on the predicted secondary structure from the first-level sequence-to-structure network (22). Most authors have since implemented this idea (in particular PSIPRED and JPred2). Pierre Baldi and colleagues deviated substantially from this concept. Instead of using an additional network, they embedded the correlation into one single recursive neural network. In principle, the idea of a recursive network had been implemented before (61). However, the particular details of the algorithm implemented in SSpro are novel and—as Table I illustrates—prove highly successful.

HMMSTR: Hidden Markov models for connecting library of structure fragments. Can we predict secondary structure for protein U by local sequence similarity to segments of known structures {S} even when overall U differs from any of the known structures {S}? Yes, as shown by many nearest-neighbor-based prediction methods, the most successful of which seems to be NSSP (62). A conceptually quite different realization of the same concept has been implemented in HMMSTR by Chris Bystroff, David Baker, and colleagues (3). First, build a library of local stretches (3–19) of residues with "basic structural motifs" (I sites). Second, assemble these local motifs through hidden Markov models introducing structural context on the level of supersecondary structure. Thus, the goal is to predict protein structure through identification of "grammatical units of protein structure formation." Although HMMSTR intrinsically aims at predicting higher order aspects of 3D structure, a side result is the prediction of 1D secondary structure. I find two results surprising. (i) The authors do not find any significant effect of "overoptimizing" their method; i.e., HMMSTR appears as accurate in predicting secondary structure

for proteins known today as it will be for those known next year. (ii) Three-state per-residue accuracy is reported to be about 74% (3). If this estimate is correct, HMMSTR is more accurate at predicting secondary structure than most existing methods and almost as accurate as the state-of-the-art methods (Table I).

And the winner is? The reason for the particular focus of this review on a small number of methods is largely that I could compare the selected methods to one another based on new proteins. A particular method that was not available to me may turn out to mark the most substantial breakthrough in the field. A Danish group developed a neural network-based method that is most amazing in many respects (63). (i) The authors estimate the method to yield levels above 77% prediction accuracy (the title of their article is slightly misleading). If true, this is the best current method. Like PSIPRED, JPred2, and PROF, the method uses PSI-BLAST profiles as input and like most methods since PHD a two-level approach addressing the problem of predicting short segments. (ii) A concept that had not been published before is to replace the standard three output units (for helix, strand, and other), by nine output units additionally coding for the secondary structure states of the residues before and after the central one (dubbed "output expansion"). (iii) Also new is the particular way of weighting the average over different networks by the overall reliability of the prediction for that network and the mere number of different networks considered (up to 800!). This impressive number of networks may prevent large-scale genome analyses based on this method. However, the major point is: Did the authors overestimate performance? The authors tested their method in a way that most developers would assume to be error-proof. However, their testing protocol is very similar to the one that I applied when significantly overestimating the accuracy of PROF (>81%). Obviously, the similarity of these two situations may very well be purely coincidental!

Plethora of new concepts for secondary structure prediction. The following five methods are a small subset of new ideas explored to improve secondary structure prediction. (i) Ouali and King (64) combine neural networks and rule-based statistics in a cascade of classifiers. Based on a similar data set they estimate a level of prediction accuracy comparable to that of JPred2 (see Table I). (ii) Chandonia and Karplus (57) combined simplified output schemes (two output states) with networks trained on different tasks and a particular variant of early stopping; input is nondivergent alignments picked from the safe zone (Fig. 1). Based on a protocol similar to that applied by the Danish group (63), the authors estimate a level of >76% accuracy, i.e., a level that if it holds up is similar to SSpro (Table I). (iii) Supposedly the simplest new method that claims to almost approach the performance of PHD combines the information for secondary structure formation contained in amino acid singlets, doublets, and triplets. (iv) Schmidler *et al.* (65) use a simple statistical model; the novel aspect is to replace compiling statistics over fixed stretches of N residues by segments signifying regular secondary structure (helix, strand). The underlying formalism resembles a hidden semi-Markov model allowing one to explicitly incorporate particular propensities such as helix caps (66). Based on noncomparable data sets the authors estimated prediction accuracy to be 69%; if correct, this is impressive for a method not using alignment information. (v) Without claims to surprising levels of accuracy, Figureau *et al.* (67) combine cleverly chosen pentapeptides from the database to obtain the final prediction.

Secondary structural class predicted almost as accurately as by experiment. Grouping proteins into secondary structure classes (all-alpha, all-beta, alpha/beta, and other) appears to be a useful initial approach for classifying proteins (27, 68). Surprisingly, such classes can be predicted successfully based merely on the overall amino acid composition of a protein (59, 69, 70). More and more increasingly complex and genial methods address this reduced goal; reported levels of prediction accuracy approach 100%. Recently, Wang and Yuan explained these high values by insufficient testing schemes and challenged that a four-state accuracy of 60% comprises the maximum for methods based solely on composition (59). Obviously, it is much easier to predict class starting from the detailed information about evolutionary profiles for the entire sequence than by restricting the input to composition. In fact, the best current methods also improve the accuracy in predicting secondary structure class considerably (Table I). The differences between observed and predicted composition of secondary structure are now below 6% for helix and strand. This is fairly close to what experimental low-resolution (circular dichroism, Fourier transform-induced spectroscopy) methods achieve at their best (57).

COMBINING MEDIOCRE AND GOOD METHODS MAY BE BEST

Combination improves on nonsystematic errors. Any prediction method has two sources of errors: (i) systematic errors, e.g., through nonlocal effects, and (ii) white noise errors caused by, e.g., the succession of the examples during training neural networks.

Theoretically, combining any number of methods improves accuracy as long as the errors of the individual methods are mutually independent and are not only systematic (71). PHD—and more recently other methods (6, 57, 63)—used this fact in combining different neural networks. The idea of combining different prediction methods has been around in secondary structure prediction for a long time (19); Cuff and Barton (see 4, 5) implemented it in JPred for different third-generation methods. In particular, JPred uses a simple expert rule for compiling the final average. King *et al.* (72) have tested a variety of different combination strategies. Selbig *et al.* (73) have compiled the jury through an elaborated decision-tree-based system. Guermeur *et al.* (74) have used a more refined variant of the JPred idea of weighting methods. Overall, combinations of independent prediction methods seem to yield levels of accuracy higher than that of the single best method. However, for every protein one method tends to be clearly superior to the combined prediction (Fig. 2B). Is it really wise to include significantly inferior methods into a combined prediction? No: averaging over all methods used for EVA decreased accuracy over the best individual methods, although averaging over the better ones was better than averaging the best ones (Rost, unpublished results). Is there any criterion for when to include a method and when not to do so? Concepts weighting the individual methods based on their accuracy and "entropy" (63) appear successful only for large numbers of methods (63; Rost, unpublished results). Nevertheless, methods that are significantly overtrained can improve when combined (Krogh, unpublished results). More rigorous studies for the optimal combination may provide a better picture. The technical problem of utilizing many methods in a public server is that the field is advancing too fast: today's methods are more accurate than averages over yesterday's methods (hence the JPred server now returns JPred2 results by default).

WHAT DOES 76% ACCURACY MEAN, IN PRACTICE?

Your protein may be predicted worse or better than average. A few problems in estimating expected prediction accuracy are described above. However, another problem is relevant for users of prediction methods: A sustained level of 76% accuracy does NOT mean that 76% of the residues in your protein of unknown structure U are correctly predicted. In contrast, prediction accuracy varies substantially between proteins (Fig. 2A). It seems that such variations are intrinsic to any method predicting aspects of protein structure and function. What can you then expect as accuracy for your protein when using a state-of-the-art method? Given a divergent family (Table II), the answer is 66–86%. Do you learn from comparing different methods?

Combining methods improves on average but you may also lose. Averaging over many methods helps, on average. However, most often some methods are more accurate than the average (Fig. 2B). Furthermore, there are examples of proteins predicted poorly by all methods (Fig. 2B), i.e., for which all methods agree by mistake (data not shown). Thus, trying to use many methods may not provide the answer to the question whether the prediction for your protein is more likely to be below or above average. Are there alternative ways to spot more reliably predicted regions?

More reliable predictions are more accurate. Reliability indices as provided by most methods correlate very well with prediction accuracy (Fig. 3). This implies that you can easily identify regions that are more likely to be predicted accurately than others. Furthermore, if your protein has many residues predicted at low levels of reliability, you may correctly suspect that your protein is predicted at a level below average. Plotting coverage versus accuracy (Fig. 3) also illustrates how beneficial more divergent profiles are to make predictions more useful. For example, PSIPRED has more than half of all residues predicted at levels that would be reached on average when comparing two known structures (75) (Fig. 3, dotted line).

ARE SECONDARY STRUCTURE PREDICTIONS USEFUL, IN PRACTICE?

Regions likely to undergo structural change predicted successfully. Young *et al.* (1) have unraveled an impressive correlation between local secondary structure predictions and global conditions. The authors monitor regions for which secondary structure prediction methods give equally strong preferences for two different states. Such regions are processed combining simple statistics and expert rules. The final method is tested on 16 proteins known to undergo structural rearrangements and on a number of other proteins. The authors report no false positives and identify most known structural switches. Subsequently, the group applied the method to the myosin family, identifying putative switching regions that were not known before, but appeared to be reasonable candidates (76). I find this method most remarkable in two ways: (i) it is the most general method using predictions of protein structure to predict some aspects of function and (ii) it illustrates that predictions may be useful even when structures are known (as in the case of the myosin family).

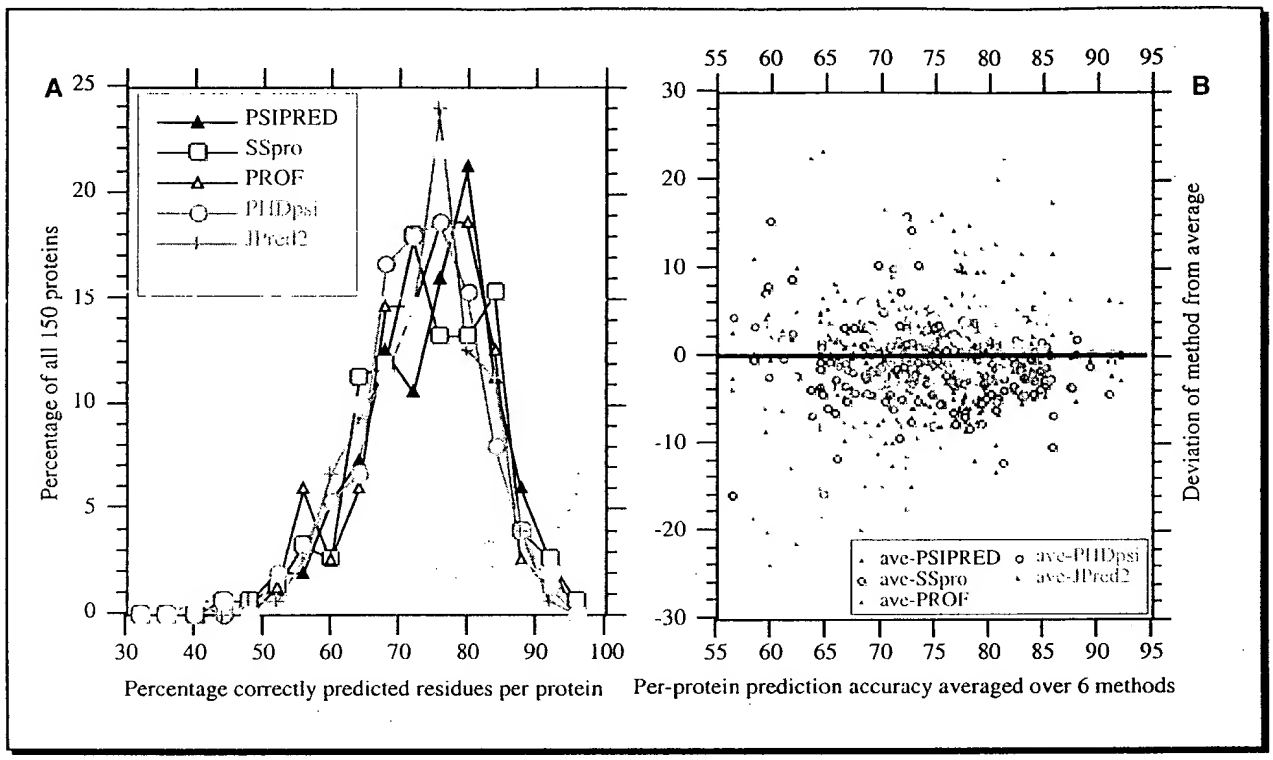

FIG. 2. Prediction accuracy varies substantially for different proteins. All results are based on 150 novel protein structures not used to develop any of the methods shown (58). The considerable difference in the three-state accuracy between different proteins is valid for all methods (A, percentage of all 150 proteins predicted at a given level of accuracy; one standard deviation is on the order of 10 percentage points). On average, different methods predict different proteins at higher levels (B, for each protein and each method, the difference between the per-protein average over all six methods is shown; negative values imply that the respective method is better than the average). *Conclusions:* (i) If you predict secondary structure for your protein with a method of 76% accuracy, the actual accuracy for that protein may be anywhere between 50 and 90%. (ii) As to be expected: most often some methods are more accurate than the average over many methods.

Classifying proteins based on secondary structure predictions in the context of genome analysis. Proteins can be classified into families based on predicted and observed secondary structure (27, 68). However, such procedures have been limited to a very coarse-grained grouping only exceptionally useful for inferring function (Table II). Nevertheless, in particular, predictions of membrane helices and coiled-coil regions are crucial for genome analysis. Recently, we came across an observation that may have important implications for structural genomics, in particular: More than one-fifth of all eukaryotic proteins appeared to have regions longer than 60 residues apparently lacking any regular secondary structure (77). Most of these regions were not of low complexity, i.e., not composition-biased. Surprisingly, these regions appeared evolutionarily as conserved as all other regions in the respective proteins. This application of secondary structure prediction may aid in classifying proteins, in separating domains, and possibly even in identifying particular functional motifs.

Aspects of protein function predicted based on expert analysis of secondary structure. The typical scenario in which secondary structure predictions facilitate learning about function is one in which experts combine their predictions and their intuition, most often to find similarities to proteins of known function but insignificant sequence similarity (39, 78–89). Usually, such applications are based on very specific details about predicted secondary structure (some examples are shown in Table II). Thus, these successful correlations of secondary structure and function appear difficult to incorporate into automatic methods.

Exploring secondary structure predictions to improve database searches. Initially, three groups independently applied secondary structure predictions for fold recognition, i.e., the detection of structural similarities between proteins of unrelated sequences (90–92). A few years later, almost every other fold recognition/threading method has adopted this concept (93–102). Two recent methods

TABLE II

Using Secondary Structure Predictions, in Practice

How to obtain the best results?	The major source of improvement is the divergence of the multiple sequence alignment used for prediction. Thus, if you have a small family, the expected prediction accuracy is lower. Particularly sensitive to divergence are the reliability indices; i.e., less divergence yields overestimated reliability indices. The most successful strategy to find the most reliably predicted regions may be to use the reliability index provided by a method rather than the agreement between different methods. If you know there are nonglobular or structural domains in your protein, chop it up before you build the alignment. If you can improve the alignment, try to do so before the prediction.
Identify membrane proteins?	Predicted membrane helices indicate that your protein is not globular. The accurate membrane predictions are usually more reliable than those for globular proteins. Thus, membrane helix predictions should be given preference. Globular methods often do not predict globular helices at positions of membrane helices; rather, often membrane helices are predicted as strand by mistakenly applied globular methods. In contrast, globular methods appear relatively more accurate for porin-like beta-strand membrane regions. Detection of membrane proteins has less than a 3% error rate for the best methods. Most helices are correctly predicted, yet the number of helices may nevertheless vary. Helix caps are clearly predicted inaccurately. Note that general methods predicting three-state secondary structure for globular proteins also predict caps less accurately.
Classify through coiled-coil regions?	Predictions of long coiled-coil regions clearly indicate that your protein is locally nonglobular. Long coiled-coil proteins are likely to be structural proteins. Longer regions are predicted more accurately.
Classify through secondary structure content?	Classifying proteins according to the secondary structure composition is helpful, but arbitrary. One hope may be to infer from the predicted secondary structure content that a particular protein is not typical. However, this attempt fails, since known protein structures vary significantly between 10 and 90% of regular secondary structure (helix, strand). Thus, secondary structure composition does not help to predict globularity.
Identify domains or structural regions?	If you see two separate secondary structure patterns, you may suspect that the protein has two structural domains. An extreme example is an N-terminal all-alpha region and a C-terminal all-beta region. If you have to cut your protein, stay more than two residues away from predicted helices and strands.
Monitor influences of point mutations?	Secondary structure prediction methods are—on average—as accurate in predicting the overall content of secondary structure as are careful CD and FTIR methods. However, such methods allow you to monitor in detail structural responses to mutations. Such changes are less likely to be reflected as accurately by prediction methods.
Find binding sites or motifs?	Most often, binding sites lie in nonregular secondary structure elements. For example, we have not predicted regular secondary structure for any of the known nuclear localization signals (128). Secondary structure predictions do not suffice to identify binding motifs, such as the zinc-finger II motif. However, the combination of sequence motif and predicted secondary structure may be very helpful.
Infer functional/structural similarity?	If you know the function/structure of protein A and want to infer whether B shares this function/structure, a similarity in the local secondary structure may help you substantially.

extended the concept by not only refining the database search, but by actually refining the quality of the alignment through an iterative procedure (50, 103). A related strategy has been implored by Ng and the Henikoffs to improve predictions and alignments for membrane proteins (104).

From 1D predictions to 2D and 3D structure. Are secondary structure predictions accurate enough to help predict higher order aspects of protein structure automatically? For 2D (interresidue contacts) predictions, Baldi *et al.* (105) have recently improved the level of accuracy in predicting β-strand

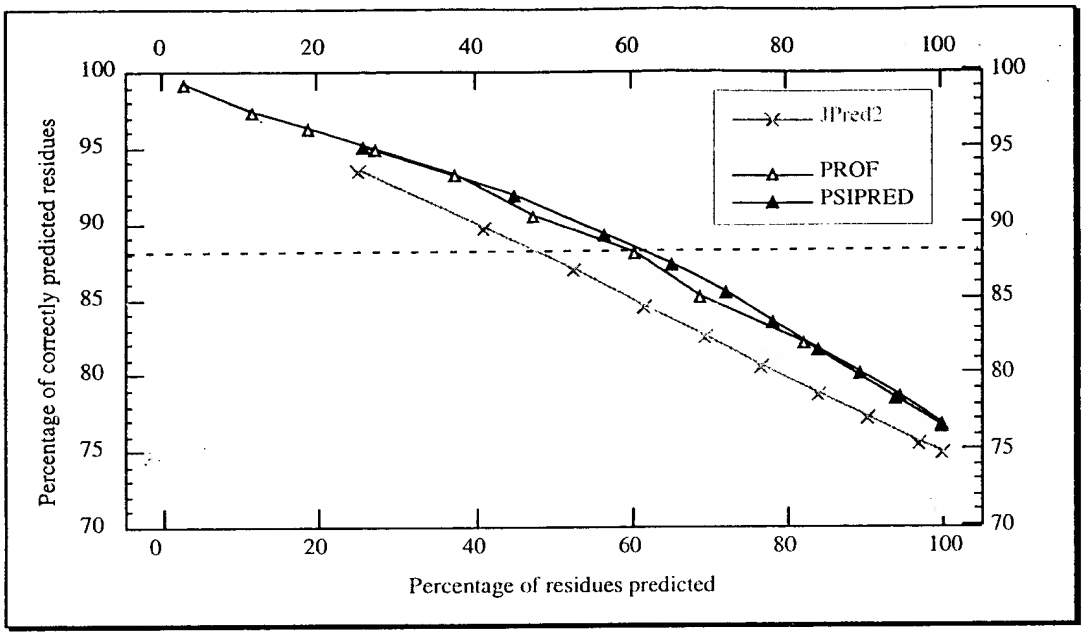

FIG. 3. Prediction accuracy correlates with reliability. The conclusion from Fig. 2A is that you have a poor idea of how well a method performs when applied to your protein of unknown structure. Fortunately, there is a way out of this dilemma: Most methods now provide an index measuring the reliability of the prediction for each residue. Shown is the accuracy versus the cumulative percentages of residues predicted at a given level of reliability (coverage vs accuracy). For example, PSIPRED and PROF reach a level above 88% for about 60% of all residues (dashed line). This particular line is chosen since secondary structure assignments by DSSP agree to about 88% for proteins of similar structure. Although JPred2 is only marginally less accurate than PSIPRED and PROF (Table I), it reaches this level of accuracy for less than half of all residues. *Conclusions:* (i) Reliability indices are extremely valuable to spot regions of more-likely-to-be-correct predictions. (ii) These indices also address the problem of variation: if many residues are predicted with high reliability, your protein is more likely to be predicted more accurately than average (Fig. 2A).

pairings over earlier work (106) by using another elaborate neural network system. For 3D predictions, the following list of five groups exemplifies that secondary structure predictions are now a popular first step toward predicting 3D structure. (i) Ortiz *et al.* (107) successfully use secondary structure predictions as one component of their 3D structure prediction method. (ii) Eyrich *et al.* (108, 109) minimize the energy of arranging predicted rigid secondary structure segments. (iii) Lomize *et al.* (110) also start from secondary structure segments. (iv) Chen *et al.* (111) suggest using secondary structure predictions to reduce the complexity of molecular dynamics simulations. (v) Levitt and co-workers (see 112, 113) combine secondary structure-based simplified presentations with a particular lattice simulation attempting to enumerate all possible folds.

AND WHAT IS THE LIMIT OF PREDICTION ACCURACY?

88% is a limit, but shall we ever reach close to there? Protein secondary structure formation is influenced by long-range interactions (45, 46, 114) and by the environment (1, 115). Consequently, stretches of up to 11 adjacent residues (dubbed chameleon after (114)) can be found in different secondary structure states (116–118). Implicitly, such nonlocal effects are contained in the exchange patterns of protein families. This is reflected by the fact that strand is predicted almost as accurately as helix (Table I), although sheets are stabilized by more nonlocal interactions than helices. Local profiles can even suffice to identify structural switches (1, 76). Surprisingly, we can find some traces of folding events in secondary structure predictions (119). Even more amazing is a study suggesting that alignment-based methods achieve levels of accuracy for chameleon regions similar to those for all other regions (118). Secondary structure assignments may vary for two versions of the same structure. One reason is that protein structures are not rocks but dynamic objects with some regions being more mobile than others. Another reason is that any assignment method must choose particular thresholds (e.g., DSSP chooses a cut-off in the Coulomb energy of a hydrogen bond). Consequently, assignments differ by about 5–15 percentage points between different X-ray versions or different NMR models for the

same protein (Andersen and Rost, unpublished results), and by about 12 percentage points between structural homologues (75). The latter number provides the upper limit for secondary structure prediction of error-free comparative modeling. I doubt that ab initio predictions of secondary structure will ever become more accurate than that. Hence, I believe a value of around 88% constitutes an operational upper limit for prediction accuracy. After the advances over the past 2 years we reached greater than 76% accuracy. Thus, we need to achieve another 12 percentage points (or even less). What is the major obstacle to reaching another 6 percentage points higher? The size of the experimental database as suggested (117)? I doubt this, since PHDpsi trained on only 200 proteins using PSI-BLAST input is almost as accurate as PSIPRED trained on 2000 proteins (Table I). Will the current explosion of sequences boost accuracy? In fact, current databases have less than 10 homologues for more than one third of the 150 proteins tested (Table I) and more than 100 for only 20% of the proteins. Although based on too a small set to draw conclusions, for these 20% highly populated families the accuracy of PROF was 4 percentage points above average (data not shown). Thus, larger databases may get us 6 percentage points higher, and it may not. The answer remains nebulous.

DISCUSSION

Methods improved significantly over the past 2 years. Growing databases and improved search techniques (Fig. 1)—predominantly through the iterated PSI-BLAST tool—yielded a substantial improvement in secondary structure prediction accuracy over the past 2 years. State-of-the-art methods now reach sustained levels of 76% prediction accuracy (Table I). Even more impressively, about 60% of all residues are predicted at levels reaching the level of agreement between X-ray and NMR structures (Fig. 3). However, novel ideas have also been shown to improve prediction accuracy. A standard way to increase the confidence in a particular prediction is to look at the results from many different prediction methods. This strategy is frequently successful and has been brought to perfection over recent years. However, often the best method is better than the average over many methods (Fig. 2B). While structure prediction is coming of age, developers and users slowly learn to reduce overestimations. However, the correlations between proteins at times of database explosions are becoming more difficult to control. It seems that only continuous, automatic evaluation servers will be able to handle this challenge in the future (58, 60).

Secondary structure predictions are at the base of structure-based sequence analysis. Almost a decade after the original breakthrough, prediction methods are now increasingly explored by wet-lab biologists to analyze their protein of interest. Secondary structure predictions are used automatically by methods aiming at higher dimensional aspects of protein structure and at improving database searches and alignment accuracy. One method has successfully related secondary structure predictions automatically to functional aspects (1, 76). However, secondary structure-based identifications of binding sites or other functional aspects are still restricted to single-case expert analyses.

And now we run human? The field has advanced considerably over the past 2 years, and more improvement appears to lie ahead. Prediction methods are fast enough to analyze entire genomes, and for particular examples the resulting classifications are relevant to structural and functional genomics (28, 68). Nevertheless, to play the devil's advocate: The field is not up to the challenge of the human sequences to be dubbed into the database very soon. We are missing a variety of approaches relating secondary structure predictions explicitly to function, such as given by ASP (1). Obviously, this remark may apply to bioinformatics, in general: The year 2001 will commence with the publication of the entire human genome; we must rush to get ready for the data flood.

Thanks are extended to Jinfeng Liu (Columbia University) for computer assistance and the collection of genome data sets; to Jinfeng Liu and Dariusz Przybylski (Columbia University) for providing preliminary information and programs; and to Claus Andersen and Søren Brunak (CBS Copenhagen) for helpful comments on the manuscript. Particular thanks are due to Volker Eyrich (Columbia University) for programming and maintaining most of the immensely valuable software that runs the EVA and META-PredictProtein servers!

REFERENCES

I find many of the publications listed here outstanding. However, I have commented on only recent publications more directly related to secondary structure prediction [except (9)]. I preferentially include comments on methods introducing new concepts and having convinced me—at least partially—that their claims hold. If the claims are true, (63) is clearly the most outstanding new development.

1. Young, M., Kirshenbaum, K., Dill, K. A., and Highsmith, S. (1999) Predicting conformational switches in proteins, *Protein Sci.* **8**, 1752–1764. Regions predicted with equally strong preferences for two secondary structure states are identified and correlated to regions undergoing structural rearrangements upon binding or environmental changes. The authors collect a data set of 16 test proteins and achieve an impressive accuracy in predicting structural switches.
2. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen

bonded and geometrical features, *Biopolymers* **22,** 2577–2637.

3. Bystroff, C., Thorsson, V., and Baker, D. (2000) HMMSTR: A hidden Markov model for local sequence–structure correlations in proteins, *J. Mol. Biol.* **301,** 173–190. The authors develop a hidden Markov model with highly branched topology to assemble local regions of protein structure. The resulting prediction of 3D structure appears often correct. Reduced to predicting secondary structure, the authors report a surprisingly high level of 74%.
4. Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. (1998) JPred: A consensus secondary structure prediction server, *Bioinformatics* **14,** 892–893.
5. Cuff, J. A., and Barton, G. J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, *Proteins* **34,** 508–519. JPred combines various prediction methods. The paper presents a good example of how to carefully evaluate prediction methods.
6. Cuff, J. A., and Barton, G. J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins* **40,** 502–511. The authors input divergent families into neural network systems. In particular, they show for the first time that PSI-BLAST alignments are competitive with dynamic programming and with hidden Markov-based alignments.
7. Rost, B. (1996) PHD: Predicting one-dimensional protein structure by profile based neural networks, *Methods Enzymol.* **266,** 525–539.
8. Przybylski, D., and Rost, B. (2000) PSI-BLAST for structure prediction: Plug-in and win, Columbia University. World Wide Web URL: http://cubic.bioc.columbia.edu/papers/.
9. Rost, B. (2000) Better secondary structure prediction through more data, Columbia University. World Wide Web URL: http://cubic.bioc.columbia.edu/predictprotein.
10. Altschul, S., Madden, T., Shaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped Blast and PSI-Blast: A new generation of protein database search programs, *Nucleic Acids Res.* **25,** 3389–3402. BLAST became useful due to its speed. PSI-BLAST extends the original concept in many ways (introducing gaps, basing alignments on position-specific profiles, allowing iterated searches, applying dynamic programming to fill regions between very similar fragments). The result is a method that is both fast and accurate. Its impact on bioinformatics continues to grow not only for secondary structure prediction.
11. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.* **292,** 195–202. David Jones was the first to successfully use PSI-BLAST profiles to automatically predict secondary structure. In this trendsetting paper, he describes some of the precautions necessary to avoid accumulating nonsimilar proteins in the iterated PSI-BLAST search.
12. Karplus, K., Barrett, C., and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies, *Bioinformatics* **14,** 846–856.
13. Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999) Exploiting the past and the future in protein secondary structure prediction, *Bioinformatics* **15,** 937–946. A fairly smart system of recursive neural network is pioneered in this work that addresses the problem of losing correlations between adjacent residues. The resulting method appears to reach levels comparable to those of PSIPRED without making full use of very diverged families.
14. Pauling, L., and Corey, R. B. (1951) Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets, *Proc. Natl. Acad. Sci. USA* **37,** 729–740.
15. Pauling, L., Corey, R. B., and Branson, H. R. (1951) The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain, *Proc. Natl. Acad. Sci. USA* **37,** 205–234.
16. Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. J., Davies, D. R., and Phillips, D. C. (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution, *Nature* **185,** 422–427.
17. Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, G., Will, G., and North, A. T. (1960) Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis, *Nature* **185,** 416–422.
18. Szent-Györgyi, A. G., and Cohen, C. (1957) Role of proline in polypeptide chain configuration of proteins, *Science* **126,** 697.
19. Rost, B., and Sander, C. (2000) Third generation prediction of secondary structure, *in* Webster, D. (Ed.), Protein Structure Prediction: Methods and Protocols, pp. 71–95, Humana Press, Clifton, NJ.
20. Dickerson, R. E., Timkovich, R., and Almassy, R. J. (1976) The cytochrome fold and the evolution of bacterial energy metabolism, *J. Mol. Biol.* **100,** 473–491.
21. Zvelebil, M. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E. (1987) Prediction of protein secondary structure and active sites using alignment of homologous sequences, *J. Mol. Biol.* **195,** 957–961.
22. Rost, B., and Sander, C. (1993) Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.* **232,** 584–599.
23. Rost, B. (1999) Twilight zone of protein sequence alignments, *Protein Eng.* **12,** 85–94.
24. Levin, J. M., Pascarella, S., Argos, P., and Garnier, J. (1993) Quantification of secondary structure prediction improvement using multiple alignment, *Protein Eng.* **6,** 849–854.
25. Barton, G. J. (1996) Protein sequence alignment and database scanning, *in* Sternberg, M. J. E. (Ed.), Protein Structure Prediction, Vol. 170, pp. 31–64, Oxford Univ. Press, London.
26. Eddy, S. R. (1998) Profile hidden Markov models, *Bioinformatics* **14,** 755–763.
27. Gerstein, M., and Levitt, M. (1997) A structural census of the current population of protein sequences, *Proc. Natl. Acad. Sci. USA* **94,** 11911–11916.
28. Teichmann, S. A., Chothia, C., and Gerstein, M. (1999) Advances in structural genomics, *Curr. Opin. Struct. Biol.* **9,** 390–399.
29. Frishman, D. (2000) PEDANT: Protein extraction, description, and analysis tool, Max-Planck-Institute, Munich. World Wide Web URL: http://pedant.mips.biochem.mpg.de/.
30. Liu, J., and Rost, B. (2000) Analysing all proteins in entire genomes, CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, World Wide Web URL: http://cubic.bioc.columbia.edu/genomes/.
31. Monne, M., Hermansson, M., and von Heijne, G. (1999) A turn propensity scale for transmembrane helices, *J. Mol. Biol.* **288,** 141–145. A particular problem of predicting membrane helices is that often short hydrophobic segments are evolutionarily conserved between membrane helices. In this paper the authors measure the propensities of forming short turns for all amino acids.

32. Lio, P., and Vannucci, M. (2000) Wavelet change-point prediction of transmembrane proteins, *Bioinformatics* **16,** 376–382.

33. Pappu, R. V., Marshall, G. R., and Ponder, J. W. (1999) A potential smoothing algorithm accurately predicts transmembrane helix packing [Published erratum appears in *Nat. Struct. Biol.* 1999, **6**(2), 199], *Nat. Struct. Biol.* **6,** 50–55.

34. Pilpel, Y., Ben-Tal, N., and Lancet, D. (1999) kPROT: A knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction, *J. Mol. Biol.* **294,** 921–935. Based on a repository of predicted membrane helices, the authors investigate the information about topology contained in the hydrophobicity scale introduced. They also show that single-span helices differ from multispan helices. Finally, the kPROT scale is shown to predict the angle of the helices with errors of less than 41°.

35. Pasquier, C., Promponas, V. J., Palaios, G. A., Hamodrakas, J. S., and Hamodrakas, S. J. (1999) A novel method for predicting transmembrane segments in proteins based on a statistical analysis of the SwissProt database: The PRED-TMR algorithm, *Protein Eng.* **12,** 381–385.

36. Chou, K. C., and Elrod, D. W. (1999) Prediction of membrane protein types and subcellular locations, *Proteins* **34,** 137–153. Membrane helices are predicted and used to distinguish between different membrane types. The work excels in many ways. However, the evaluation of prediction accuracy is not fully convincing.

37. Kühlbrandt, W., and Gouaux, E. (1999) Membrane proteins, *Curr. Opin. Struct. Biol.* **9,** 445–447.

38. Rost, B., and O'Donoghue, S. I. (1997) Sisyphus and prediction of protein structure, *Comput. Appl. Biosci.* **13,** 345–356.

39. de Fays, K., Tibor, A., Lambert, C., Vinals, C., Denoel, P., De Bolle, X., Wouters, J., Letesson, J. J., and Depiereux, E. (1999) Structure and function prediction of the *Brucella abortus* P39 protein by comparative modeling with marginal sequence similarities, *Protein Eng.* **12,** 217–223.

40. Lupas, A. (1997) Predicting coiled-coil regions in proteins, *Curr. Opin. Struct. Biol.* **7,** 388–393.

41. Wolf, E., Kim, P. S., and Berger, B. (1997) MultiCoil: A program for predicting two- and three-stranded coiled coils, *Protein Sci.* **6,** 1179–1189.

42. O'Donoghue, S. I., and Nilges, M. (1997) Tertiary structure prediction using mean-force potentials and internal energy functions: Successful prediction for coiled-coil geometries, *Folding Design* **2,** S47–S52.

43. Kolinski, A., Skolnick, J., Godzik, A., and Hu, W. P. (1997) A method for the prediction of surface "U"-turns and transglobular connections in small proteins, *Proteins* **27,** 290–308.

44. Shepherd, A. J., Gorse, D., and Thornton, J. M. (1999) Prediction of the location and type of beta-turns in proteins using neural networks, *Protein Sci.* **8,** 1045–1055.

45. Muñoz, V., Cronet, P., López-Hernández, E., and Serrano, L. (1996) Analysis of the effect of local interactions on protein stability, *Folding Design* **1,** 167–178.

46. Villegas, V., Zurdo, J., Filimonov, V. V., Aviles, F. X., Dobson, C. M., and Serrano, L. (2000) Protein engineering as a strategy to avoid formation of amyloid fibrils, *Protein Sci.* **9,** 1700–1708.

47. Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., and Sillitoe, I. (1999) Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction, *Proteins* **37,** 149–170.

48. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L., and Hughey, R. (1999) Predicting protein structure using only sequence information, *Proteins* **S3,** 121–125.

49. Cuff, J. A., Birney, E., Clamp, M. E., and Barton, G. J. (2000) ProtEST: Protein multiple sequence alignments from expressed sequence tags, *Bioinformatics* **16,** 111–116.

50. Jennings, A. J., Edge, C. M., and Sternberg, M. J. E. An approach to improve multiple alignments of protein sequences using predicted secondary structure, *Protein Eng.*, in press. The authors use secondary structure predictions to improve alignment accuracy. For a few examples, the method is shown to be beneficial.

51. Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996) Using CLUSTAL for multiple sequence alignments, *Methods Enzymol.* **266,** 383–402.

52. King, R. D., and Sternberg, M. J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, *Protein Sci.* **5,** 2298–2310.

53. Altschul, S. F., and Gish, W. (1996) Local alignment statistics, *Methods Enzymol.* **266,** 460–480.

54. Bairoch, A., and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, *Nucleic Acids Res.* **28,** 45–48.

55. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res.* **28,** 235–242.

56. Rost, B., and Sander, C. (1994) 1D secondary structure prediction through evolutionary profiles, *in* Bohr, H., and Brunak, S. (Eds.), Protein Structure by Distance Analysis, pp. 257–276, IOS Press, Amsterdam.

57. Chandonia, J. M., and Karplus, M. (1999) New methods for accurate prediction of protein secondary structure, *Proteins* **35,** 293–306.

58. Rost, B., Eyrich, V. A., Przybylski, D., Pazos, F., Valencia, A., Fiser, A., Marti-Renom, M., Sanchez, R., and Sali, A. (2000) EVA—Evaluation of automatic protein structure prediction services, Columbia University/Rockefeller University/CNB Madrid. World Wide Web URL: http://cubic.bioc.columbia.edu/eva.

59. Wang, Z.-X., and Yuan, Z. (2000) How good is prediction of protein structural class by the component-coupled method? *Proteins* **38,** 165–175. The authors argue that most methods predicting secondary structural class from amino acid composition have significantly overestimated performance accuracy. They suggest that approaches based on composition alone can never reach above 60%. The method they develop is estimated at slightly above 50% accuracy.

60. Rychlewski, W. W. W. L., and Fischer, D. (2000) LiveBench: Continuous benchmarking of prediction servers, IIMCB Warsaw. World Wide Web URL: http://BioInfo.PL/LiveBench/.

61. Reczko, M. (1993) Protein secondary structure prediction with partially recurrent neural networks, *in* First International Workshop on Neural Networks Applied to Chemistry and Environmental Sciences, Lyon, France, pp. 153–159, Gordon and Breach, New York.

62. Salamov, A. A., and Solovyev, V. V. (1997) Protein secondary structure prediction using local alignments, *J. Mol. Biol.* **268,** 31–36.

63. Petersen, T. N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G. P., and Lund, O. (2000) Prediction of protein secondary structure at 80% accuracy, *Proteins* **41,** 17–20. The authors use divergent PSI-BLAST profiles to train and test neural networks. Novel is the particular way of averaging over many networks, as well as the amazing number of networks averaged (up to 800). The authors also replace the standard three output units (helix, strand, and other) by nine units coding for the three secondary structure states of three adjacent residues. Prediction accuracy is estimated to be higher than 77%.
64. Ouali, M., and King, R. D. (2000) Cascaded multiple classifiers for secondary structure prediction, *Protein Sci.* **9,** 1162–1176.
65. Schmidler, S. C., Liu, J. S., and Brutlag, D. L. (2000) Bayesian segmentation of protein secondary structure, *J. Comput. Biol.* **7,** 233–248.
66. Aurora, R., and Rose, G. D. (1998) Helix capping, *Protein Sci.* **7,** 21–38.
67. Figureau, A., Angelica Soto, M., and Toha, J. (1999) Secondary structure of proteins and three-dimensional pattern recognition, *J. Theor. Biol.* **201,** 103–111.
68. Przytycka, T., Aurora, R., and Rose, G. D. (1999) A protein taxonomy based on secondary structure, *Nat. Struct. Biol.* **6,** 672–682.
69. Liu, W., and Chou, K. C. (1999) Prediction of protein secondary structure content, *Protein Eng.* **12,** 1041–1050.
70. Zhang, C. T., and Zhang, R. (1999) Skewed distribution of protein secondary structure contents over the conformational triangle, *Protein Eng.* **12,** 807–810.
71. Hansen, L. K., and Salamon, P. (1990) Neural network ensembles, *IEEE Pattern Anal. Machine Intell.* **12,** 993–1001.
72. King, R. D., Ouali, M., Strong, A. T., Aly, A., Elmaghraby, A., Kantardzic, M., and Page, D. (2000) Is it better to combine predictions? *Protein Eng.* **13,** 15–19.
73. Selbig, J., Mevissen, T., and Lengauer, T. (1999) Decision tree-based formation of consensus protein secondary structure prediction, *Bioinformatics* **15,** 1039–1046.
74. Guermeur, Y., Geourjon, C., Gallinari, P., and Deleage, G. (1999) Improved performance in protein secondary structure prediction by inhomogeneous score combination, *Bioinformatics* **15,** 413–421.
75. Rost, B., Sander, C., and Schneider, R. (1994) Redefining the goals of protein secondary structure prediction, *J. Mol. Biol.* **235,** 13–26.
76. Kirshenbaum, K., Young, M., and Highsmith, S. (1999) Predicting allosteric switches in myosins, *Protein Sci.* **8,** 1806–1815.
77. Liu, T. J., Tan, H., and Rost, B. (2000) Genomes full of proteins with long non-structured regions? Columbia University. World Wide Web URL: http://cubic.bioc.columbia.edu/papers/.
78. Paquet, J. Y., Vinals, C., Wouters, J., Letesson, J. J., and Depiereux, E. (2000) Topology prediction of *Brucella abortus* Omp2b and Omp2a porins after critical assessment of transmembrane beta strands prediction by several secondary structure prediction methods, *J. Biomol. Struct. Dyn.* **17,** 747–757.
79. Di Stasio, E., Sciandra, F., Maras, B., Di Tommaso, F., Petrucci, T. C., Giardina, B., and Brancaccio, A. (1999) Structural and functional analysis of the N-terminal extracellular region of beta-dystroglycan, *Biochem. Biophys. Res. Commun.* **266,** 274–278.
80. Juan, H. F., Hung, C. C., Wang, K. T., and Chiou, S. H. (1999) Comparison of three classes of snake neurotoxins by homology modeling and computer simulation graphics, *Biochem. Biophys. Res. Commun.* **257,** 500–510.
81. Laval, V., Chabannes, M., Carriere, M., Canut, H., Barre, A., Rouge, P., Pont-Lezica, R., and Galaud, J. (1999) A family of *Arabidopsis* plasma membrane receptors presenting animal beta-integrin domains, *Biochimica et Biophysica Acta* **1435,** 61–70.
82. Seto, M. H., Liu, H. L., Zajchowski, D. A., and Whitlow, M. (1999) Protein fold analysis of the B30.2-like domain, *Proteins* **35,** 235–249.
83. Xu, H., Aurora, R., Rose, G. D., and White, R. H. (1999) Identifying two ancient enzymes in Archaea using predicted secondary structure alignment, *Nat. Struct. Biol.* **6,** 750–754.
84. Jackson, R. M., and Russell, R. B. (2000) The serine protease inhibitor canonical loop conformation: Examples found in extracellular hydrolases, toxins, cytokines and viral proteins, *J. Mol. Biol.* **296,** 325–334.
85. Stawiski, E. W., Baucom, A. E., Lohr, S. C., and Gregoret, L. M. (2000) Predicting protein function from structure: Unique structural features of proteases, *Proc. Natl. Acad. Sci. USA* **97,** 3954–3958.
86. Shah, P. S., Bizik, F., Dukor, R. K., and Qasba, P. K. (2000) Active site studies of bovine alpha1→3-galactosyltransferase and its secondary structure prediction, *Biochimica et Biophysica Acta* **1480,** 222–234.
87. Brautigam, C., Steenbergen-Spanjers, G. C., Hoffmann, G. F., Dionisi-Vici, C., van den Heuvel, L. P., Smeitink, J. A., and Wevers, R. A. (1999) Biochemical and molecular genetic characteristics of the severe form of tyrosine hydroxylase deficiency, *Clin. Chem.* **45,** 2073–2078.
88. Davies, G. P., Martin, I., Sturrock, S. S., Cronshaw, A., Murray, N. E., and Dryden, D. T. (1999) On the structure and operation of type I DNA restriction enzymes, *J. Mol. Biol.* **290,** 565–579.
89. Gerloff, D. L., Cannarozzi, G. M., Joachimiak, M., Cohen, F. E., Schreiber, D., and Benner, S. A. (1999) Evolutionary, mechanistic, and predictive analyses of the hydroxymethyldihydropterin pyrophosphokinase family of proteins, *Biochem. Biophys. Res. Commun.* **254,** 70–76.
90. Fischer, D., and Eisenberg, D. (1996) Fold recognition using sequence-derived properties, *Protein Sci.* **5,** 947–955.
91. Russell, R. B., Copley, R. R., and Barton, G. J. (1996) Protein fold recognition by mapping predicted secondary structures, *J. Mol. Biol.* **259,** 349–365.
92. Rost, B. (1995) TOPITS: Threading one-dimensional predictions into three-dimensional structures, *in* Rawlings, C., Clark, D., Altman, R., Hunter, L., Lengauer, T., and Wodak, S. (Eds.), Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, England, pp. 314–321, AAAI Press, Menlo Park, CA.
93. de la Cruz, X., and Thornton, J. M. (1999) Factors limiting the performance of prediction-based fold recognition methods, *Protein Sci.* **8,** 750–759.
94. Di Francesco, V., Munson, P. J., and Garnier, J. (1999) FORESST: Fold recognition from secondary structure predictions of proteins, *Bioinformatics* **15,** 131–140.
95. Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM, *J. Mol. Biol.* **299,** 499–520.
96. Ayers, D. J., Gooley, P. R., Widmer-Cooper, A., and Torda, A. E. (1999) Enhanced protein fold recognition using second-

ary structure information from NMR, *Protein Sci.* **8,** 1127–1133.

97. Hargbo, J., and Elofsson, A. (1999) Hidden Markov models that use predicted secondary structures for fold recognition, *Proteins* **36,** 68–76.
98. Jones, D. T. (1999) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences, *J. Mol. Biol.* **287,** 797–815.
99. Panchenko, A., Marchler-Bauer, A., and Bryant, S. H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence, *Proteins* **Suppl. 3,** 133–140.
100. Ota, M., Kawabata, T., Kinjo, A. R., and Nishikawa, K. (1999) Cooperative approach for the protein fold recognition, *Proteins* **37,** 126–132.
101. Koretke, K. K., Russell, R. B., Copley, R. R., and Lupas, A. N. (1999) Fold recognition using sequence and secondary structure information, *Proteins* **37,** 141–148.
102. Jones, D. T., Tress, M., Bryson, K., and Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure, *Proteins* **37,** 104–111.
103. Heringa, J. (1999) Two strategies for sequence comparison: Profile-preprocessed and secondary structure-induced multiple alignment, *Comput. Chem.* **23,** 341–364. Alignment consistency is checked and alignments are improved through preprocessing the profile and using predicted secondary structure. The resulting method is shown to yield more sensitive database searches for a few examples.
104. Ng, P., Henikoff, J., and Henikoff, S. (2000) PHAT: A transmembrane-specific substitution matrix, *Bioinformatics* **16,** 760–766. A membrane-specific exchange matrix is collected. Then alignments of membrane proteins are refined by an iterative procedure using and improving predictions of membrane helices.
105. Baldi, P., Pollastri, G., Andersen, C. A., and Brunak, S. (2000) Matching protein beta-sheet partners by feedforward and recurrent neural networks, *Ismb* **8,** 25–36.
106. Hubbard, T. J. P., and Park, J. (1995) Fold recognition and ab initio structure predictions using hidden Markov models and β-strand pair potentials, *Proteins* **23,** 398–402.
107. Ortiz, A. R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., and Skolnick, J. (1999) Ab initio folding of proteins using restraints derived from evolutionary information, *Proteins* **Suppl. 3,** 177–185.
108. Eyrich, V. A., Standley, D. M., Felts, A. K., and Friesner, R. A. (1999) Protein tertiary structure prediction using a branch and bound algorithm, *Proteins* **35,** 41–57.
109. Eyrich, V. A., Standley, D. M., and Friesner, R. A. (1999) Prediction of protein tertiary structure to low resolution: Performance for a large and structurally diverse test set, *J. Mol. Biol.* **288,** 725–742.
110. Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I. (1999) Prediction of protein structure: The problem of fold multiplicity, *Proteins* **Suppl.,** 199–203.
111. Chen, C. C., Singh, J. P., and Altman, R. B. (1999) Using imperfect secondary structure predictions to improve molecular structure computations, *Bioinformatics* **15,** 53–65.
112. Samudrala, R., Xia, Y., Huang, E., and Levitt, M. (1999) Ab initio protein structure prediction using a combined hierarchical approach, *Proteins* **Suppl.,** 194–198.
113. Samudrala, R., Huang, E. S., Koehl, P., and Levitt, M. (2000) Constructing side chains on near-native main chains for ab initio protein structure prediction, *Protein Eng.* **13,** 453–457.
114. Minor, D. L. J., and Kim, P. S. (1996) Context-dependent secondary structure formation of a designed protein sequence, *Nature* **380,** 730–734.
115. Krittanai, C., and Johnson, W. C. J. (2000) The relative order of helical propensity of amino acids changes with solvent environment, *Proteins* **39,** 132–141.
116. Zhou, X., Alber, F., Folkers, G., Gonnet, G. H., and Chelvanayagam, G. (2000) An analysis of the helix-to-strand transition between peptides with identical sequence, *Proteins* **41,** 248–256.
117. Pan, X. M., Niu, W. D., and Wang, Z. X. (1999) What is the minimum number of residues to determine the secondary structural state? *J. Protein Chem.* **18,** 579–584.
118. Jacoboni, I., Martelli, P. L., Fariselli, P., Compiani, M., and Casadio, R. (2000) Predictions of protein segments with the same amino acid sequence and different secondary structure: A benchmark for predictive methods, *Proteins* **41,** 535–544. Some stretches of up to 11 adjacent residues are known to adopt different secondary structure in different structural contexts (chameleon regions). In this original work, the authors show that, surprisingly, profile-based neural network predictions are almost as accurate for such chameleon regions as they are for regions that are never observed in alternative states.
119. Compiani, M., Fariselli, P., Martelli, P. L., and Casadio, R. (1999) Neural networks to study invariant features of protein folding, *Theor. Chem. Accounts* **101,** 21–26. The authors successfully identify likely nucleation sites, as well as which helix forms first from secondary structure predictions.
120. Abagyan, R. A., and Batalov, S. (1997) Do aligned sequences share the same fold? *J. Mol. Biol.* **273,** 355–368.
121. Park, J., Teichmann, S. A., Hubbard, T., and Chothia, C. (1997) Intermediate sequences increase the detection of distant sequence homologies, *J. Mol. Biol.* **273,** 349–354.
122. Rost, B., and Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure, *Proteins* **19,** 55–72.
123. Zemla, A., Venclovas, C., Fidelis, K., and Rost, B. (1999) A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment, *Proteins* **34,** 220–223.
124. Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta* **405,** 442–451.
125. Kneller, D. G., Cohen, F. E., and Langridge, R. (1990) Improvements in protein secondary structure prediction by an enhanced neural network, *J. Mol. Biol.* **214,** 171–182.
126. Zhang, C.-T., and Chou, K.-C. (1992) An optimization approach to predicting protein structural class from amino acid composition, *Protein Sci.* **1,** 401–408.
127. Defay, T., and Cohen, F. E. (1995) Evaluation of current techniques for ab initio protein structure prediction, *Proteins* **23,** 431–445.
128. Cokol, M., Nair, R., and Rost, B. (2001) Finding nuclear localisation signals, *EMBO Rep.* **1(5),** 411–415.